Journal of KIIT. Vol. 23, No. 7, pp. 149-154, Jul. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.7.149

AI 음성 인식 향상을 위한 개선된 엔트로피 특징 추출

오 상 엽*

Improvement Entropy Feature Extraction for AI Voice Recognition Improvement

Sang Yeob Oh*

요 약

최근 인공 지능 기반 음성 처리 기술은 딥 러닝을 기반으로 한 CNN, DNN, RNN, 그리고 Kaldi 등의 방법으로 음성에 대한 인식률은 향상되었다. 그러나 이와 같은 기술을 적용한 인공 지능 방법에서도 비슷한 발음의 단어나 잡음으로 인한 단어의 누락, 연음 법칙으로 인한 음성 오류 등이 발생되는 문제를 가진다. 이와 같은 문제를 해결하기 위해 본 연구에서는 음성 특징 추출에서 음성과 비음성 에너지 스펙트럼의 엔트로피 특징 추출을 이용하여 잡음 환경에서 음성을 정확히 처리하기 위한 모델을 제안한다. 제안된 방법의 실험 분석에서는 SNR(Signal to Noise Ratio) 변화에 의한 음성 인식 성능을 다양한 잡음 환경에서 실험하여 제안된 방법의 실험에서 음성 인식의 성능 향상이 개선된 것을 확인하였다.

Abstract

Recently, artificial intelligence-based speech processing technology has improved the recognition rate of speech with methods such as CNN, DNN, RNN, TDNN, and Kaldi based on deep learning. However, artificial intelligence methods that apply such technologies also have problems such as omission of words with similar pronunciations or words due to noise, and speech errors due to the law of linking sounds. To solve this problem, this paper proposes a model to accurately process speech in a noisy environment by using entropy feature extraction of speech and non-speech energy spectra in speech feature extraction. In the experimental analysis of the proposed method, the speech recognition performance according to the change in Signal to Noise Ratio (SNR) was tested in various noisy environments, and it was checked that the performance of speech recognition was improved in the experiment of the proposed method.

Keywords
vocabulary recognition, feature extraction, SNR, noise elimination, energy spectrum

* 가천대학교 컴퓨터공학과 교수
- ORCID: ttps://orcid.org//0000-0002-8002-9588

· Received: Apr. 09, 2025, Revised: Jun. 11, 2025, Accepted: Jun. 14, 2025
· Corresponding Author: Sang Yeob Oh
Dept. of Computer Engineering, Gachon University, Korea
Tel.: +82-31-750-5798, Email: syoh@gachon.ac.kr

Ⅰ. 서 론

음성 인식은 AI 기술의 발전으로 인해 네비게이션, 통역 등의 분야에서 발전하여 음성 감정 인식 등의 분야로 발전하고 있다[1][2]. 대부분의 음성 인식 기반의 데이터는 영어를 기반으로 연구되고 있으며, 한국어에서의 음성 인식 처리는 비슷한 발음의 다른 음성 처리, 유사 발음의 다른 음성 인식 문제, 잡음으로 인한 음성 누락, 연음 법칙으로 인한 음성 인식 오류 등이 발생할 수 있다[3].

이와 같은 한국어 음성 인식의 특성상 가장 주요한 음성 데이터의 품질에 영향을 주는 문제는 정확한 특징 추출과 음성 데이터 주변의 잡음이다. 기존의 음성 특징 추출을 위한 방법으로는 MFCC(Mel Frequency Cepstral Coefficient)가 있으며, 최근에는 이 기반으로 추출된 특징 벡터에 데이터 처리를 위한 RNN(Recurrent Neural Network)과 음성 데이터의 시간적 패턴과 주파수 특징을 학습할 수 있는 CNN(Convolution Neural Network) 모델을 사용한다[4]. 이와 같은 학습 모델의 적용을 위해서는 학습되는 음성 데이터가 중요한 문제로 대두되고 있다.

또한, 음성 인식에서는 다양한 환경 조건에서 잡음이 발생하므로 잡음에 대해 정확하게 인식하는 문제가 중요하다. 음성 신호와 비슷한 신호 특징을 갖거나 시간 변화에 따라 통계적인 신호 특성이 변화하는 불안정한 잡음은 제거율이 떨어지는 문제점을 가지고 있다[5]. 그러므로 음성 신호 처리 시스템에서 음성 인식 성능 분석의 정확도에 있어서 잡음에 대한 SNR 분석은 음성인식의 성능과 정확도에 중요하다. 음성 인식 처리 시스템에서 다양한 잡음을 최소화하기 위한 방법과 이를 기반으로 음성의 정확도를 높이기 위한 개선 방법이 제안되었다[6]-[8].

또한, 잘못된 잡음의 추정에 따라 음성 신호에 존재하는 잡음이 유지되어 음성 신호에 해당하는 스펙트럼이 변형되거나 음성 신호에 해당하는 프레임을 검색하지 못하고 음성 인식 성능을 저하시키는 문제가 발생된다.

이와 같은 문제는 음성 인식을 위한 모델 훈련 환경과 음성 인식 환경의 차이에서 발생되며, 이러한 환경의 차이와 함께 높은 SNR에서는 음성(Voiced)과 비음성(Unvoiced)에 대한 특징을 추출하기 위한 인식 모델을 제안하며, 이 방법에서 음성과 비음성에 대한 에너지 스펙트럼의 엔트로피 특징 추출을 이용하여 잡음 환경에서 음성을 정확히 처리하기 위한 방안을 제시한다. 이를 위해 엔트로피는 DFT(Discrete Fourier Transform)에 대해 개선된 에너지 스펙트럼 엔트로피를 적용하여 잡음 환경에 강인한 음성 특징을 검출한다. 따라서 음성의 특징이 잡음의 영향을 적게 받을 수 있도록 모델을 구성하였고, SNR의 분석을 위해 SDR(Signal to Distortion ratio)을 이용하여 음성과 비음성의 Cepstrum 특징 분포 특성을 이용하였으며, 인식률 확인 결과 기존 방법에 비해 향상된 인식률을 확인하였다.

본 논문은 2장 관련 연구에서는 SNR과 특징 추출에 대한 내용을 다루며, 3장의 시스템 모델에서는 엔트로피 음성 신호 특징 추출을 이용한 잡음 환경에 강인한 방법을 제안하였다. 4장 성능 분석에서는 음성에 대한 평가를 SDR을 이용하여 다른 논문과 비교 및 평가하였으며, 그리고 5장은 결론으로 구성한다.

Ⅱ. 관련 연구

2.1 SNR

SNR은 음성 인식에 대한 신호에서 소음, 잡음, 간섭 등의 잡음 비율을 나타내며, 음성 신호와 와 잡음 사이의 상대적인 비율 강도를 표현하며, 음성 3신호와 잡음에 대한 상대적인 세기를 비교하는 데 이용된다. 일반적으로 SNR은 잡음 대비 음성 신호의 비율을 가지고 상대적인 음성 신호 분포 정도를 나타내며, 음성 처리 시스템의 성능이 절대적인 음성 신호 전력이 아닌 잡음 대비 신호로 결정되며, 다음 식 (1)로 표현된다[9].

$$SNR = 10^{*}\log\left(\frac{P_{signal}}{P_{noise}}\right) \quad (1)$$

Pnoise는 잡음의 크기, Psignal은 음성 신호를 나타내며, 잡음이 신호에 대한 영향을 정량적으로 나타내는 척도로 사용된다. 일반적으로 높은 SNR은 좋은 음질이나 정확한 음성인식을 나타낸다.

음성 인식에 포함된 잡음은 다양하게 발생되며, [9][10], 발생된 잡음을 제거하기 위해서는 음성 데이터를 표본화한 모델을 사용하거 학습된 임펄스 응답 필터를 이용하며[11], 응답 필터는 음성 데이터의 사용에 의해 부가되는 잡음을 삭제하는 기능을 가진다. 음성 인식 처리에서 음성에 혼합된 잡음을 디지털 데이터로 처리하기 위해 임펄스 응답과 선형 시변 필터를 적용하여 처리한다.

$$y[n]=\sum_{l=0}^{+\infty} f[n,l]s[n-1]+noise[n] \quad (2)$$

식 (2)에서 음성 입력 신호 $s[n]$과 임펄스 응답 $f[n,l]$, 그리고 출력 $y[n]$을 의미하며, 음성에 대한 임펄스 응답 $f[l]$에 대한 $f[n,l]$에서 l은 1보다 크거나 같다.

2.2 특징 추출

특징 추출(Feature extraction)은 음성 인식 데이터에서 음성에 대한 데이터 신호로 부터 추출하며, 이를 처리하기 위해 음성 데이터의 압축, 데이터의 단순화 과정으로 추출된 대표 특성들을 가지고 인식률을 판단한다. 특징 추출 과정에서 사용되는 기본적인 방법은 청각 특성을 반영하는 달팽이관의 주파수 응답을 이용하여 필터 뱅크 분석을 적용한다. 음성 인식 신호의 시간의 변화에 의한 변환 특성을 처리하기 위하여 셉스트럼(Cepstrum)에 대한 1차와 2차 미분 값을 적용한다. 미분 값은 시간 축 방향의 필터링으로 표현된다. 시간의 변화에 의한 음성 데이터의 특징 벡터를 얻기 위해 주로 사용되는 방법에는 MFCC(Mel Frequency Cepstrum Coefficient), 그리고 LPC(Linear Predictive Coefficient)가 사용된다. MFCC에서는 음성 신호에서 발생되는 잡음을 anti aliasing 필터를 이용하여 아날로그. 데이터를 디지털 데이터로 처리한다. MFCC 에서는 데이터 처리를 위해 프레임 로그 에너지를 부가하여 음성 데이터의 특성 벡터로 이용된다. LPC는 신호의 스펙트럼 및 FFT Cepstrum으로 얻은 값을 음성 데이터 특징 추출에 사용한다. 또한, 음성 신호에 부가된 잡음을 최소화하기 위해 음성 데이터의 잡음을 최소화 한 특징을 추출하는 방법으로는 mel-cepstrum을 사용한다[8], 이 방법에서는 band-pass 필터를 사용하여 음성을 여러 개의 필터 뱅크에 처리하였으며, 각 음성에 대한 에너지를 가지고 음성에 대한 특징을 추출한다[10].

III. 시스템 모델

음성 신호는 폐의 압력으로부터 나오는 공기가 성대와 구강을 거쳐 나오면서 진동과 흐름에 의해 음성과 비음성으로 구분된다. 공기의 진동과 흐름은 신호의 세기를 가지면서 음성 신호를 에너지 양으로 측정하게 된다. 에너지 양이 크면 음성구간으로, 작으면 비음성 구간으로 분류하여 dB로 측정이 가능하다. 그림 1에서 에너지 스펙트럼을 갖는 엔트로피를 에너지 스펙트럼으로 표현하였다. 음성은 비음성에 비해 비교적 낮은 주파수대에 나타나는 주기적 신호이며, 비음성은 음성에 비해 높은 주파수대에 나타나는 비주기적 신호 특징을 가진다.

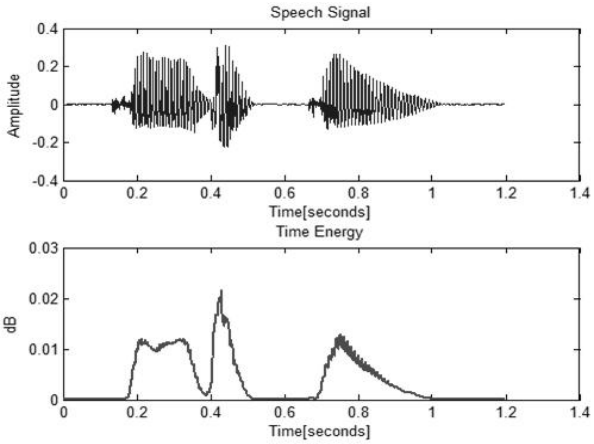

그림 1. 음성 신호와 시간 에너지
Fig. 1. Speech signal and time energy

이와 같은 신호의 특징을 이용하여 음성과 비음성 선택 신호에 의하여 음성과 비음성에 적합한 음성 소스를 선택적으로 적용하고, 에너지 양이 크게 나타나는 저주파 구간인 음성구간에서 ZCR(Zero

Crossing Rate) 값이 낮은 결과 값을 가지며, 비음성 구간에서는 ZCR 값이 높은 결과 값을 보인다. 식 (3)에서 음성구간에 대한 time domain의 각 샘플을 지정하여 정해진 각 구간에 대해 제곱하고 가산된 값을 누적하여 Log-energy로 사용하며, 수식은 다음과 같다.

$$\log E = 10\log[\sum_{k=1}^{n}|S_k^2|] \tag{3}$$

음성 특징은 음성 신호의 기본 주파수를 나타내는 것으로 펄스 음성 소스의 대표 구간을 특징 주기마다 복원하고, 펄스 음성 소스 데이터를 생성하여 프레임 마다 펄스 특징을 추출한다. 에너지 스펙트럼은 엔트로피와 동일하게 처리하므로 스펙트럼 에너지 영역을 엔트로피로 정의하여 사용한다. 엔트로피는 DFT(Discrete Fourier Transform)를 이용하여 계산하며 수식은 다음과 같다.

$$P(|Y(k,l)|^2) = \frac{|Y(k,l)|^2}{\sum_{k=1}^{N/2}|Y(k,l)|^2} \tag{4}$$

식 (4)에서 k는 frequency bin에 대한 인덱스 값을 의미하며, l은 프레임 인덱스를 의미한다. 프레임에서 frequency bin에 대한 스펙트랄 에너지 확률을 계산하고, 처리된 각 frequency bin의 확률은 엔트로피로 표현한다. 개선된 에너지 스펙트럼 엔트로피 확률 수식은 다음과 같으며, 에너지 스펙트럼을 갖는 엔트로피를 표현하였다.

$$S(|Y(k,l)|^2) = -\sum_{k=1}^{N/2} P(Y(k,l)|^2 \cdot \log_2(P(|Y(k,l)|^2)) \tag{5}$$

잡음 제거는 여러 채널들의 입력 신호를 입력 받은 기준 신호나 실험적 결과 값과 신호 대 잡음비 등을 이용하여 추정되어진 잡음 신호를 사용한다. 이 경우에 여러 개의 채널이 필요하거나 높은 연산의 양이 필요하므로 분해된 음성신호를 음성 밴드와 잡음밴드로 분리한다. 분리된 잡음 신호 값을 가지고 프레임에서 음성 구간의 각 밴드에서 발생하는 계수의 편차 값을 가지고 특징 벡터를 처리한다. 한글 초성 자음은 상대적으로 신호가 매우 약하기 때문에 초성 자음이 음성으로 검출되지 않는 경우가 있다. 이는 초성을 제외한 오류로 인식률에 큰 영향을 주므로 이러한 문제를 해결하고자 식 (5)에서 log를 사용하여 초성 자음 부분의 에너지를 크게 하여 정확한 음성으로 검출되도록 한다.

Ⅳ. 성능 분석

본 연구에서는 음성과 비음성에 대한 에너지 스펙트럼의 엔트로피 특징 추출을 DFT를 적용하여 잡음 환경에 강인한 음성 특징을 검출하는 방법을 제안하였으며, 성능에 대한 평가 수행은 음성 향상 algorithm의 성능 검증으로 사용되는 Aurora 2.0을 사용하였다[5][6][10]. 음성 데이터베이스 Aurora 2.0에서 SNR 변화에 따른 음성 인식 성능을 실험하였다. Aurora 2.0은 여러 잡음 환경과 레벨에 대한 데이터 집합을 가지고 백색 가우시안 잡음과 혼합 잡음에 대한 음성 신호 성능 향상 검증으로 사용된다. 음성 신호에 대한 잡음 신호 처리를 수행하기 위해 8kHz sampling rate, 16bit를 사용하고, FFT 크기는 256 샘플을 사용하며, 신호 변형을 감소하기 의해 해밍 원도(Hamming window)를 이용하였다. 음성 신호 비교 분석에서 잡음 처리를 위해 워너 필터를 사용하고, 음성 인식 실험을 위해 서울 지역명과 지하철역명에 대해 각각 20개를 사용하였다. 음성에 대한 인식을 수행 평가하기 위해 음성 인식 내용을 각 5회 발음하여 20개의 지역 명에 대해 100단어를 사용하였으며, 음성에 대한 평가는 SDR[5][6]을 이용하여 수행하였다.

$$SDR(dB) = 10\log_{10}\frac{\sum_{n=1}^{N}[x(n)-\hat{x}(n)]^2}{\sum_{n=1}^{N}x^2(n)} \tag{6}$$

식 (6)에서 $x(n)$는 잡음이 없는 음성을 나타내고, $\hat{x}(n)$는 잡음이 부가된 음성 신호를 나타내며, n은 시간 인덱스를 의미한다.

본 연구의 비교 분석을 위해 클러스터링 모델과

비모수 상관 계수를 이용한 방법[5]과 개선된 MFCC와 가우시안 잡음 제거를 적용한 방법[6]을 비교하여 성능 평가를 수행한 결과를 표 1에 나타내었다.

표 1에서의 분석 결과 같이 식 (6)의 값이 작으면 비교된 음성의 잡음에 대한 인식률을 효율적으로 처리하였음을 나타낸다. 비교 논문 [5]에 비해서는 인식률이 0.22dB이 개선되었으며, 비교 논문 [6]에 대해서는 0.03dB 개선되었으며, 제안된 방법이 분리 음성 60%에서 음성에 대한 인식률이 [6] 에 비해 개선되었음을 나타낸다.

표 1. 결과 분석
Table 1. Compare with the results

Separated voice	Paper results of [5]	Paper results of [6]	Suggested method
1	2.13	2.10	2.09
2	1.51	1.31	1.37
3	1.99	1.79	1.71
4	1.67	1.67	1.66
5	1.79	1.39	1.41
6	2.38	2.38	2.31
7	1.83	1.93	1.85
8	6.14	5.17	5.31
9	2.16	2.06	2.08
10	2.60	2.51	2.31
11	1.59	1.59	1.60
12	2.29	2.27	2.28
13	2.50	2.53	2.51
14	2.14	2.14	2.07
15	2.11	2.19	2.14
16	2.79	2.71	2.31
17	2.11	2.11	2.12
18	2.77	2.51	2.67
19	2.00	2.31	2.27
20	2.10	2.14	2.11
Voice average	2.33	2.24	2.21

V. 결 론

한국어 음성 인식의 특성상 품질에 영향을 주는 문제는 정확한 특징 추출과 음성 데이터 주변의 잡음이다. 본 연구에서는 음성과 비음성에 대한 에너지 스펙트럼의 엔트로피 특징 추출을 이용하여 잡음 환경에서 음성을 정확히 처리하기 위해 엔트로피는 DFT(Discrete Fourier Transform)에 대해 개선된 에너지 스펙트럼 엔트로피를 적용하여 잡음 환경에 강인한 음성 특징을 검출하여 음성의 특징이 잡음의 영향을 적게 받을 수 있도록 모델을 구성하였고, SNR의 분석을 위해 SDR(Signal to Distortion ratio)을 이용하여 음성과 비음성의 Cepstrum 특징 분포 특성을 이용하여 인식률을 향상시켰다. 이러한 결과는 인공지능 기반 환경에서의 음성 인식을 보다 향상시키는데 기여할 수 있을 것으로 기대된다.

References

[1] M. H. Yi and J. H. Shin, "Emotion Recognition Model Using Progressive Transfer Learning for Speech Data Adation", Jounal of Korea Multimedia Society, Vol. 27, No. 8, pp. 1004-1013, Aug. 2024.

[2] Y.-J. kim, H.-J. Cha, and A R. Kang, "A Study on the Impact of Speech Data Quality on Speech Recognition Model", Journal of the Korea Society of Computer and information, Vol. 29, No. 1, pp. 41-29, Jan. 2024. https://doi.org/10.9708/jksci.2024.29.01.041.

[3] H. Jin, A.-H. Lee, Y.-J. Chae, S.-H. Park, Y.-J. Kang, and S. W. Lee, "Error correction for Korean Speech Recognition using a LSTM-based Sequence to Sequence Model", Journal of the Korea Society of Computer and information, Vol. 26, No. 10, pp. 1-7, Oct. 2021. https://doi.org/10.9708/jksci.2021.26.10.001.

[4] E. D. Cahyadi, H. N. H. Soesild, and M. Song, "Enhancing Multimodal Emotion Recogniтion in Speech and Text with Integrated CNN, LSTM, and BERT Models", The Journal of Convergence on Culture Technology, Vol. 10, No. 1, pp. 617-623, Jan. 2024. https://doi.org/10.17703/JCCT.2024.10.1.617.

[5] S.-Y. Oh, "Vocabulary Recognition Rate

Enhancement using Clustering Model and Non-parametric Correlation Coefficient", The Journal of Korean Institute of Information Technology, Vol. 22, No. 4, pp. 91-97, Apr. 2024. https://doi.org/10.14801/jkiit.2024.22.4.91.

[6] S.-Y. Oh, "Noise Elimination Using Improved MFCC and Gaussian Noise Deviation Estimation", Journal of The Korea Society of Computer and Information Vol. 28 No. 1, pp. 87-92, Jan. 2023. https://doi.org/10.9708/jksci.2023.28.01.087.

[7] S.-Y. Oh, "Speech Recognition Performance Improvement using a convergence of GMM Phoneme Unit parameter and Vocabulary Clustering", Journal of Convergence for Information Technology, Vol. 10, No. 8, pp. 35-39, Aug. 2020. https://doi.org/10.22156/CS4SMB.2020.10.08.035.

[8] S.-Y. Oh, "DNN based Robust Speech Feature Extraction and Signal Noise Removal Method Using Improved Average Prediction LMS Filter for Speech Recognition", Journal of Convergence for Information Technology, Vol. 11, No. 6, pp. 1-6, Jun. 2021. https://doi.org/10.22156/CS4SMB.2021.11.06.001.

[9] D. H. Johnson "Signal-to-Noise Ratio", Scholar pedia, Vol. 1, No. 12, pp. 2088, 2006.

[10] K. Chung and S. Y. Oh, "Vocabulary optimization process using similar phoneme recognition and feature extraction", Cluster Computing, Vol. 19, No. 3, pp. 1683-1690, Aug. 2016. https://doi.org/10.1007/s10586-016-619-0.

[11] K. Chung and S. Y. Oh, "Voice Activity Detection Using an Improved Unvoiced Feature Normalization Process in Noisy Environments", Wirekess Personal Communications, Vol. 89, No. 3, pp. 747-759, 2016. https://doi.org/10.1007/s11277-015-3169-5.

저자소개

오 상 엽(Sang Yeob Oh)

1989년 2월 : 경원대학교 전자계산학과(공학사)
1991년 2월 : 광운대학교 전자계산학과(이학석사)
1998년 2월 : 광운대학교 전자계산학과(이학박사)
1992년 9월 ~ 현재 : 가천대학교 컴퓨터공학과 교수
관심분야 : 음성 인식, 잡음 검출, 음성 특징 추출, 멀티미디어 데이터 통신